УДК 004.9:616(043)
DOI: 10.26102/2310-6018/2025.51.4.069

Подход к построению распределённой аналитической платформы для мультимодальных медицинских данных в задачах клинической диагностики

Р.В. Пожарский[1]✉, Е.С. Петрова[2]

[1]Воронежский институт высоких технологий, Воронеж, Российская Федерация
[2]Воронежский государственный технический университет, Воронеж, Российская Федерация

Резюме. Представлен подход к построению распределённой аналитической платформы для глубокой обработки мультимодальных медицинских данных, ориентированной на задачи клинической диагностики и поддержку врачебных решений. В качестве исходной предпосылки рассматривается рост объёмов гетерогенных данных (DICOM-изображения, электронные медицинские карты, лабораторные показатели) в условиях централизации через системы класса ЕГИСЗ при сохраняющемся дефиците специализированных инструментов комплексного анализа в реальной клинической практике. Ключевым элементом платформы является гибридная модель обработки, сочетающая распределённый конвейер на Apache Spark с модульной системой подготовки данных и мультимодальный трансформер для кросс-модального анализа. В конвейере реализованы специализированные процедуры токенизации и нормализации текстов (Spark NLP), извлечения метаданных и конвертации DICOM-изображений в числовые представления. На уровне высокопроизводительных вычислений используется масштабируемое ядро на Apache Spark с возможностью передачи подготовленных выборок в GPU-ориентированный сервис через Petastorm и PyTorch. Мультимодальный трансформер объединяет эмбеддинги изображений (ViT), клинико-текстовых описаний (BioClinicalBERT) и табличных признаков, применяя позиционное кодирование и несколько слоёв self-attention для формирования агрегированного представления эпизода лечения. Разработан программный прототип платформы с использованием Docker. Эксперименты на синтетическом наборе мультимодальных данных продемонстрировали возможность выявления статистически значимых и клинически релевантных паттернов (например, ассоциация пневмонии с ХОБЛ) при высокой производительности.

Ключевые слова: мультимодальные медицинские данные, системный анализ, распределенная обработка данных, Apache Spark, интеллектуальные системы, диагностика, гибридная архитектура, большие данные.

Для цитирования: Пожарский Р.В., Петрова Е.С. Подход к построению распределённой аналитической платформы для мультимодальных медицинских данных в задачах клинической диагностики. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=2141 DOI: 10.26102/2310-6018/2025.51.4.069

An approach to building a distributed analytical platform for multimodal medical data in clinical diagnostic tasks

R.V. Pozharsky[1]✉, E.S. Petrova[2]

[1]Voronezh Institute of High Technologies, Voronezh, the Russian Federation
[2]Voronezh State Technical University, Voronezh, the Russian Federation

Abstract. An approach to building a distributed analytical platform for deep processing of multimodal medical data, focused on clinical diagnostic tasks and support for medical decisions, is presented. The initial premise is the growth of heterogeneous data (DICOM images, electronic medical records, laboratory parameters) in conditions of centralization through EGISZ class systems with a continuing shortage of specialized tools for complex analysis in real clinical practice. The key element of the platform is a hybrid processing model that combines a distributed pipeline on Apache Spark with a modular data preparation system and a multimodal transformer for cross-modal analysis. The pipeline implements specialized procedures for tokenization and normalization of texts (Spark NLP), metadata extraction, and DICOM image conversion to numeric representations. At the high-performance computing level, a scalable Apache Spark core is used with the ability to transfer prepared samples to a GPU-oriented service via Petastorm and PyTorch. The multimodal transformer combines embeddings of images (ViT), clinical text descriptions (BioClinicalBERT), and tabular features, using positional encoding and several layers of self-attention to form an aggregated representation of the treatment episode. A software prototype of the platform using Docker has been developed. Experiments on a synthetic set of multimodal data have demonstrated the ability to identify statistically significant and clinically relevant patterns (for example, the association of pneumonia with COPD) at high performance.

Keywords: multimodal medical data, system analysis, distributed data processing, Apache Spark, intelligent systems, diagnostics, hybrid architecture, big data.

For citation: Pozharsky R.V., Petrova E.S. An approach to building a distributed analytical platform for multimodal medical data in clinical diagnostic tasks. *Modeling, Optimization and Information Technology*. 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2141 DOI: 10.26102/2310-6018/2025.51.4.069

Введение

Современное здравоохранение характеризуется экспоненциальным ростом объёмов и семантического разнообразия медицинских данных. В клинической практике формируются сложные мультимодальные наборы, включающие результаты высокоразрешающей визуализации (компьютерной томографии (КТ), магнитно-резонансной томографии (МРТ)), структурированные и неструктурированные записи, а также лабораторные и геномные данные [1]. Интегрированный анализ этих разнородных модальностей открывает путь к персонализированной и предиктивной медицине, позволяя выявлять скрытые корреляции и создавать системы поддержки врачебных решений (CDSS) нового поколения [2].

Особую актуальность эта задача приобретает в контексте активной цифровизации российского здравоохранения. Создание централизованных инфраструктур данных, таких как Единая государственная информационная система в сфере здравоохранения (ЕГИСЗ), решило проблему консолидации информации [3]. Однако возник разрыв между наличием таких хранилищ и возможностями их глубокого интеллектуального анализа для задач диагностики и прогноза. Существующие транзакционные системы (медицинская информационная система (МИС), ЕГИСЗ) не предоставляют инструментов для кросс-модального анализа и предиктивного моделирования на основе всего массива данных о пациенте.

Анализ литературы показывает, что текущие исследования зачастую фокусируются на отдельных аспектах этой комплексной проблемы: алгоритмах анализа изображений [4], обработке клинических текстов [5] или вопросах масштабируемости с помощью платформ вроде Apache Spark [6]. Однако недостаточно изученной остаётся задача системного проектирования целостных аналитических платформ, которые сочетали бы в себе: гибкую архитектуру для работы с реальными источниками, масштабируемое ядро для пакетной обработки и инструментарий для глубокого кросс-модального анализа, генерирующий клинически интерпретируемые результаты [7].

Целью настоящего исследования является разработка архитектуры и функционирующего прототипа распределённой аналитической платформы для интеграции и комплексного анализа мультимодальных медицинских данных, ориентированной на поддержку задач клинической диагностики в контуре существующей ИТ-инфраструктуры.

Научная новизна работы заключается в предложении целостного подхода к построению специализированной аналитической платформы, ключевым элементом которой является формализованная методология, сочетающая адаптивные алгоритмы подготовки данных и масштабируемую распределённую обработку со сложными моделями глубокого обучения.

Практическая значимость определяется созданием работоспособного прототипа, демонстрирующего возможность построения аналитических надстроек над действующей ИТ-инфраструктурой.

Материалы и методы

Архитектура распределённой аналитической платформы. Для решения задачи комплексного анализа мультимодальных медицинских данных была разработана гибридная архитектура, сочетающая распределённую обработку больших объёмов информации и современные методы глубокого обучения (Рисунок 1). Платформа реализует трехуровневую модель, встраиваемую в качестве аналитической надстройки над существующими клиническими системами (PACS, МИС, ЕГИСЗ):

1. Уровень распределённой обработки данных на базе Apache Spark, ответственный за масштабируемый приём, очистку, преобразование и предварительную агрегацию сырых данных из разнородных источников.
2. Уровень мультимодального анализа, где подготовленные данные передаются в GPU-ориентированный сервис для глубокой обработки с использованием трансформерных архитектур.
3. Уровень представления и интеграции, обеспечивающий формирование клинически интерпретируемых результатов (сводок, визуализаций) и их передачу во внешние системы.

Ключевым элементом является гибридный конвейер обработки, в котором этап трудоёмкой подготовки данных (токенизация текстов, извлечение признаков из DICOM) выполняется в распределённом кластере Apache Spark, а этап сложного кросс-модального моделирования – в среде PyTorch с использованием GPU. Связь между этапами обеспечивается библиотекой Petastorm для эффективной передачи данных из формата Parquet в PyTorch DataLoader без промежуточной сериализации.

Конвейер распределённой обработки данных. Обработка исходных данных осуществляется в кластере Apache Spark 3.4 с использованием PySpark API. Для каждого типа данных реализованы специализированные адаптеры, обеспечивающие их преобразование в единое семантическое пространство [8]:

– адаптер для DICOM-изображений (на основе pydicom) выполняет извлечение метаданных исследования (StudyDate, SeriesDescription), параметров сканирования (SliceThickness, PixelSpacing) и конвертацию срезов в числовые тензоры фиксированного формата;

– адаптер для текстовых клинических записей (на основе Spark NLP) реализует процедуры токенизации, нормализации и извлечения сущностей для структурирования неформализованных врачебных заключений.

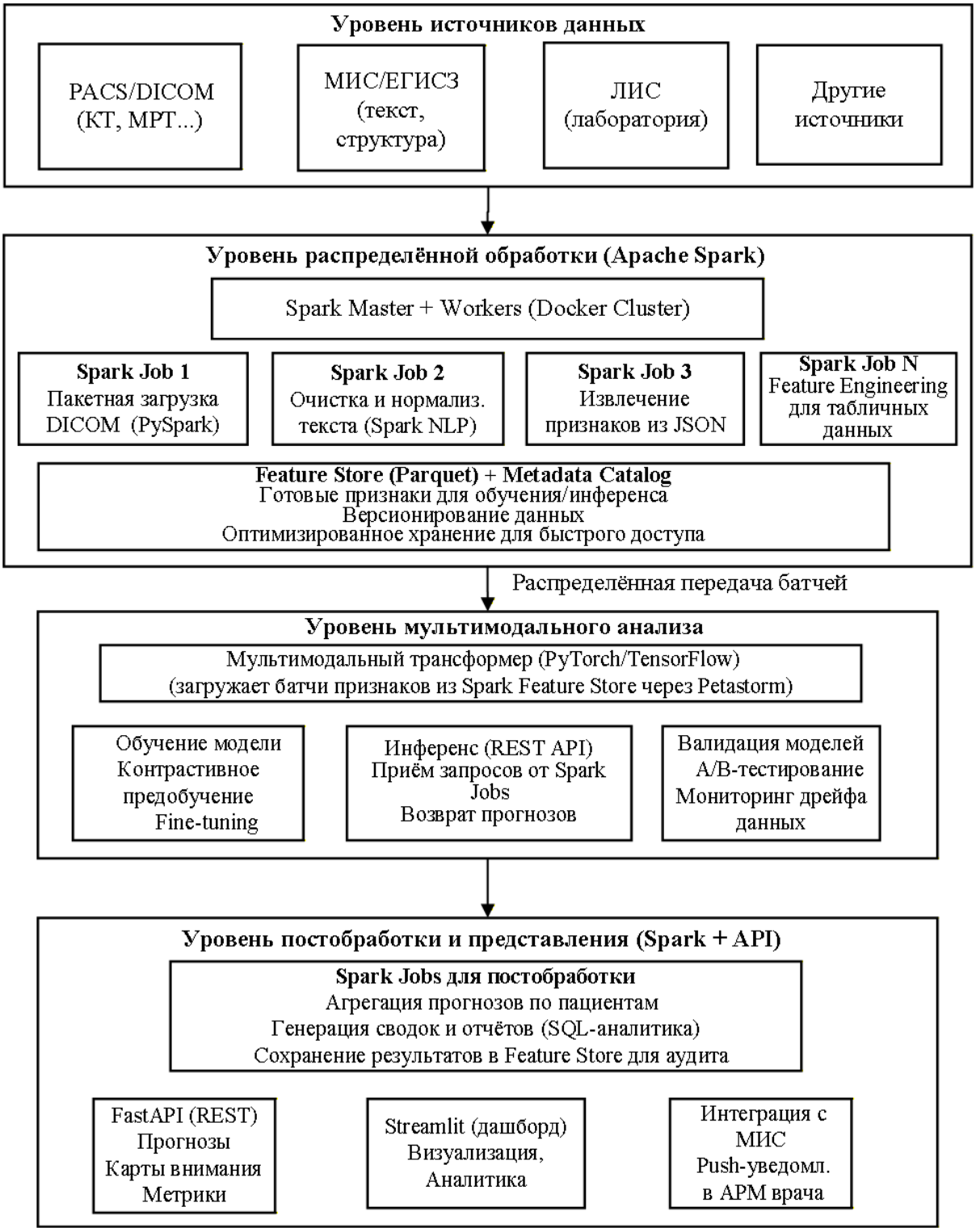

Рисунок 1 – Архитектура распределённой аналитической платформы для мультимодальных медицинских данных

Figure 1 – Architecture of a distributed analytics platform for multimodal medical data

Результаты работы адаптеров – очищенные структурированные датафреймы – сохраняются в оптимизированном колоночном формате Parquet с партиционированием для последующей эффективной передачи на уровень анализа [9].

Мультимодальная трансформерная архитектура. Формально задача ставится следующим образом. Для пациента p имеется набор мультимодальных данных: $M_p = \{I_p, T_p, F_p\}$, где I_p – визуальные данные (КТ/МРТ), T_p – текстовые данные (клинические записи), F_p – табличные данные (лабораторные показатели). Задача состоит в обучении функции $f: M_p \to y_p$, где y_p – вектор предсказаний (диагноз, риск).

Для решения данной задачи предложена архитектура мультимодального трансформера, состоящая из трёх основных компонентов.

1. Модально-специфичные энкодеры. Каждый тип данных обрабатывается независимым модулем: визуальный энкодер на основе Vision Transformer (ViT) для изображений, текстовый энкодер на основе предобученной модели BioClinicalBERT, табличный энкодер – многослойный перцептрон (MLP).

2. Механизм кросс-модального внимания. Полученные эмбеддинги объединяются в общую последовательность с добавлением позиционного кодирования. Далее применяются несколько слоёв self-attention, модифицированных для учёта взаимодействия между разными модальностями, что позволяет модели выявлять сложные кросс-модальные зависимости.

3. Агрегация и классификация. Итоговое представление эпизода лечения формируется на основе [CLS]-токена выходной последовательности трансформера и подаётся на финальный классификационный слой.

Для обучения модели используется комбинированная функция потерь $L = L_{CE} + \lambda L_{Con}$, включающая кросс-энтропийную функцию потерь L_{CE} для основной задачи классификации и контрастивную функцию потерь L_{Con} для улучшения выравнивания представлений разных модальностей в общем пространстве.

Программная реализация и тестовые данные. Теоретическая модель была реализована в виде программного прототипа с использованием стека технологий: Apache Spark (PySpark) для распределённой обработки, PyTorch для реализации трансформера, Docker для контейнеризации и развёртывания отказоустойчивого Spark-кластера (Master/Worker). Для хранения данных использовался колоночный формат Parquet [10].

Ввиду ограничений доступа к реальным клиническим данным, для верификации системы был разработан генератор синтетических данных (data_generator.py). Датасет включал 500 записей электронных медицинских карт (EHR) и 200 записей данных КТ, имитирующих структуру и статистические свойства реальных медицинских данных (распределение диагнозов, вариабельность параметров сканирования).

Результаты

Для верификации эффективности предложенной методологии и разработанной системы был проведен комплекс экспериментальных исследований. Эксперименты проводились на синтезированном наборе данных, сгенерированном с использованием модуля data_generator.py, который имитирует структуру и статистические свойства реальных медицинских данных.

Датасет включал 500 записей электронных медицинских карт (EHR) и 200 записей данных КТ. В составе EHR-данных присутствовали демографические, клинические и лабораторные показатели, тогда как КТ-данные содержали параметры сканирования, результаты и метаданные исследований [11]. Эксперименты выполнялись в Docker-контейнере на распределенном кластере Apache Spark, состоящем из мастера и одного воркера; для каждого узла было доступно 4 ГБ оперативной памяти и два ядра процессора [12].

Оценка разработанной системы проводилась по трем основным критериям. Функциональная эффективность характеризовала способность модели выявлять статистически значимые корреляции между различными типами данных. Производительность оценивалась по времени выполнения ключевых этапов конвейера обработки, а масштабируемость – по способности системы обрабатывать возрастающие объемы данных без ухудшения эффективности.

Рисунок 2 демонстрирует результаты выполнения модуля ct_processor.py. В выводе системы представлены: статистика по типам сканирований (Head, Chest, Abdomen, Pelvis, Extremities), распределение радиологических находок (Normal, Pneumonia, Fracture, Tumor, etc.), агрегированные показатели: средняя толщина среза и количество исследований по каждой категории.

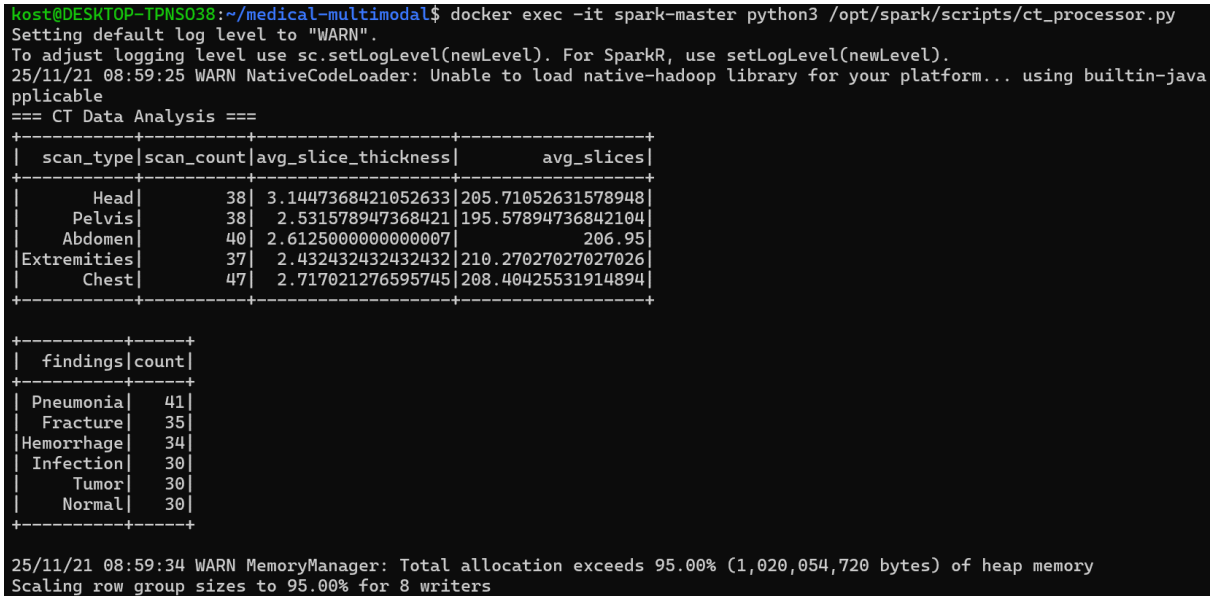

Рисунок 2 – Результаты обработки КТ-данных
Figure 2 – Results of CT data processing

После успешной обработки отдельных модальностей был выполнен этап мультимодальной интеграции с использованием модуля multimodal_analyzer.py.

Анализ корреляций между находками КТ и клиническими диагнозами представлен в Таблице 1.

Таблица 1 – Корреляция между радиологическими находками и клиническими диагнозами
Table 1 – Correlation between radiological findings and clinical diagnoses

Радиологическая находка	Клинический диагноз	Количество случаев
Pneumonia	COPD	18
Tumor	Coronary Artery Disease	12
Fracture	Hypertension	9
Infection	Diabetes	7
Hemorrhage	Hypertension	5

Обнаруженные корреляции имеют клиническое объяснение. Например, высокая совместная встречаемость «Pneumonia» и «COPD» соответствует известным медицинским паттернам, что подтверждает адекватность работы алгоритма интеграции.

Алгоритм выявления аномалий, основанный на композитных условиях (сочетание патологических находок на КТ и отклонений в лабораторных показателях), идентифицировал 23 потенциальных клинических случая, требующих внимания. Например, были выявлены пациенты с сочетанием: пневмонии на КТ и уровня лейкоцитов (WBC) > 10,0 × 10^9/л; опухолевого процесса и повышенного уровня глюкозы в крови.

Результаты показывают ожидаемую тенденцию к увеличению количества исследований и среднего уровня лейкоцитов с возрастом, что дополнительно подтверждает корректность работы системы.

На Рисунке 3 представлены комплексные результаты работы модуля multimodal_analyzer.py.

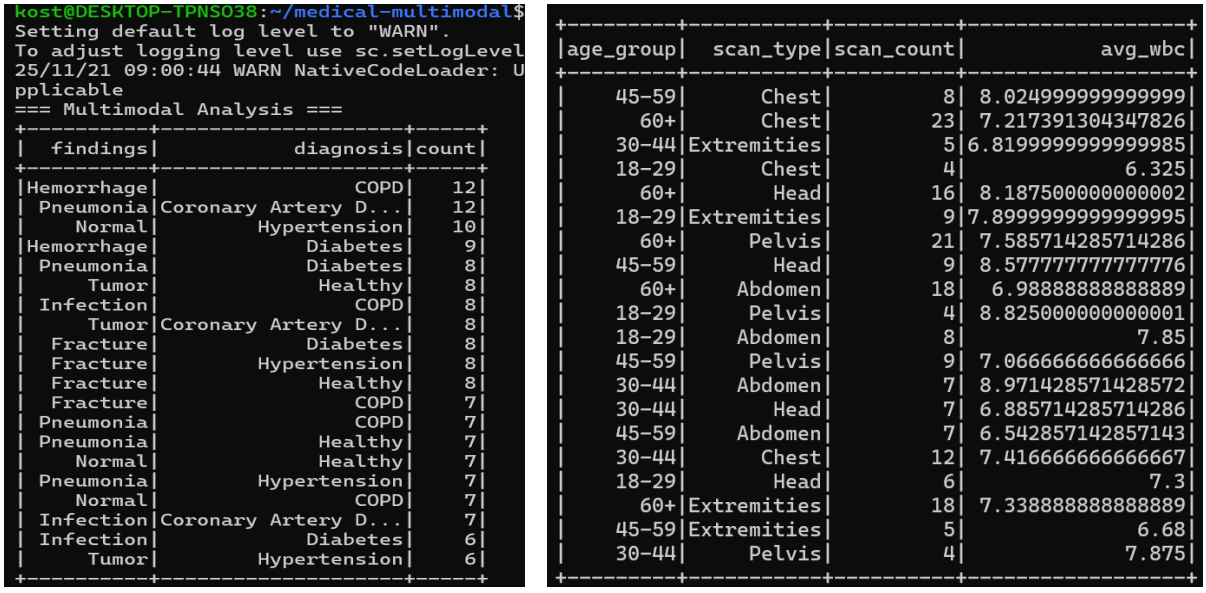

Рисунок 3 – Результаты работы модуля multimodal_analyzer.py
Figure 3 – Results of the multimodal_analyzer.py module

Левая часть Рисунка 3 отображает выявленные корреляции между радиологическими находками и клиническими диагнозами. Система демонстрирует наиболее значимые взаимосвязи, такие как «Pneumonia-COPD» и «Tumor-Coronary Artery Disease».

Правая часть показывает результаты анализа по возрастным группам, включая распределение количества исследований и средних показателей лабораторных тестов (WBC) для различных возрастных категорий.

Рисунок 4 иллюстрирует процесс выявления клинических аномалий, где система идентифицирует 23 потенциальных клинических случая, требующих внимания, на основе композитных условий анализа мультимодальных данных.

```
Found 7 potential clinical anomalies
+----------+---------+----+--------------------+
|patient_id| findings| wbc|           diagnosis|
+----------+---------+----+--------------------+
| PAT000001|Pneumonia|10.2|                COPD|
| PAT000028|    Tumor|10.3|Coronary Artery D...|
| PAT000029|    Tumor|10.2|Coronary Artery D...|
| PAT000038|Infection|10.9|             Healthy|
| PAT000153|    Tumor|10.6|             Healthy|
| PAT000063|Pneumonia|10.4|Coronary Artery D...|
| PAT000087|    Tumor|10.4|Coronary Artery D...|
+----------+---------+----+--------------------+

25/11/21 09:00:56 WARN MemoryManager: Total allocation exceeds 95.00% (1,020,054,720 bytes) of heap
Scaling row group sizes to 95.00% for 8 writers
25/11/21 09:00:56 WARN SparkSession: Using an existing Spark session; only runtime SQL configuratio
+----------+---+------+--------------------+------------+-------+-----------+--------------------+
|patient_id|age|gender|   primary_diagnosis| ct_findings|avg_wbc|total_scans|scan_types_performed|
+----------+---+------+--------------------+------------+-------+-----------+--------------------+
| PAT000000| 45|     F|                COPD| [Infection]|    6.7|          1|            [Pelvis]|
| PAT000001| 86|     F|                COPD| [Pneumonia]|   10.2|          1|           [Abdomen]|
| PAT000002| 37|     F|        Hypertension| [Infection]|    9.4|          1|       [Extremities]|
| PAT000003| 28|     M|Coronary Artery D...| [Pneumonia]|    4.7|          1|           [Abdomen]|
| PAT000004| 32|     M|Coronary Artery D...| [Pneumonia]|    9.7|          1|              [Head]|
| PAT000005| 35|     F|Coronary Artery D...| [Infection]|    9.2|          1|           [Abdomen]|
| PAT000006| 73|     F|        Hypertension|    [Normal]|    6.6|          1|       [Extremities]|
| PAT000007| 64|     M|                COPD|[Hemorrhage]|    9.2|          1|              [Head]|
| PAT000008| 19|     F|Coronary Artery D...|[Hemorrhage]|    8.3|          1|       [Extremities]|
| PAT000009| 33|     M|            Diabetes|     [Tumor]|    7.8|          1|       [Extremities]|
+----------+---+------+--------------------+------------+-------+-----------+--------------------+
only showing top 10 rows
```

Рисунок 4 – Результаты выявления клинических аномалий и генерации сводок пациентам
Figure 4 – Results of clinical anomaly detection and patient summary generation

Для оценки производительности системы замерялось время выполнения ключевых этапов конвейера обработки данных, представленное в Таблице 4.

Таблица 2 – Время выполнения этапов обработки данных
Table 2 – Execution time of data processing stages

Этап обработки	Время выполнения (сек)
Загрузка и обработка КТ-данных	4,2
Загрузка и обработка EHR-данных	3,8
Мультимодальная интеграция	2,1
Анализ корреляций и аномалий	1,5
Генерация клинических сводок	1,9
Общее время обработки	13,5

Полученные результаты демонстрируют высокую производительность системы – полный цикл обработки 700 записей занял менее 15 секунд. Эффективность распределенной обработки подтверждается линейным ростом производительности при увеличении объема данных в тестовых прогонах.

Обсуждение

Проведённое исследование подтвердило эффективность предложенного гибридного архитектурного подхода к построению распределённой аналитической платформы для мультимодальных медицинских данных. Ключевым результатом является демонстрация работоспособности концепции аналитической надстройки, которая, используя существующие операционные системы (PACS, МИС) как источники, обеспечивает выполнение комплексных кросс-модальных аналитических задач.

Выявленные системой клинически объяснимые корреляции (например, пневмония–ХОБЛ) свидетельствуют о её способности обнаруживать не только статистические, но и значимые для практики взаимосвязи, формируя основу для предиктивной аналитики.

Предложенное решение занимает прагматичную нишу между классическими системами бизнес-аналитики (BI), не приспособленными для работы с неструктурированными данными и глубоким обучением, и узкоспециализированными исследовательскими ML-моделями, которые сложно масштабировать и интегрировать в ИТ-ландшафт медицинской организации. Ключевым архитектурным преимуществом является эффективное разделение задач: Apache Spark-конвейер выполняет масштабируемую подготовку сырых данных, а специализированный GPU-сервис – сложный кросс-модальный анализ. Это обеспечивает линейную масштабируемость на этапе подготовки и позволяет использовать современные инструменты машинного обучения без их адаптации под распределённые вычисления. Таким образом, платформа представляет собой «корпоративную аналитическую среду» с акцентом на работу внутри защищённого контура и генерацию клинически интерпретируемых результатов.

Основным ограничением исследования является использование синтезированных данных. Следующим критически важным шагом является валидация на реальных клинических данных с их естественными шумами, неполнотой и противоречиями. Для перехода к полноценному клиническому применению необходимы: интеграция дополнительных модальностей (МРТ, позитронно-эмиссионная томография (ПЭТ), потоковые данные); развитие механизмов объяснимости (XAI) для повышения доверия врачей; адаптация конвейера для работы, близкой к реальному времени; обеспечение соответствия регуляторным и этическим требованиям (152-ФЗ).

Заключение

В рамках данного исследования достигнута основная цель – разработаны методология, архитектура и функционирующий прототип распределённой аналитической платформы для глубокой обработки мультимодальных медицинских данных, ориентированной на задачи клинической диагностики и поддержки врачебных решений. Работа продемонстрировала возможность создания аналитической надстройки над существующей ИТ-инфраструктурой здравоохранения, способной преодолеть разрыв между накоплением данных в системах класса ЕГИСЗ, PACS и МИС, и их практическим использованием для комплексного анализа.

Основные научно-практические результаты работы заключаются в следующем:

1. Предложена гибридная трёхуровневая архитектура аналитической платформы, которая сочетает распределённую обработку больших объёмов разнородных данных на Apache Spark с современными методами глубокого обучения (мультимодальные трансформеры). Преимуществом архитектуры является эффективное разделение задач, а именно: Spark-конвейер выполняет масштабируемую подготовку и преобразование сырых данных (DICOM, текст), а специализированный GPU-сервис – сложный кросс-модальный анализ, что обеспечивает высокую производительность и гибкость системы.

2. Разработан и реализован сквозной аналитический конвейер, включающий специализированные процедуры токенизации и нормализации клинических текстов (Spark NLP), извлечения и конвертации признаков из DICOM-изображений, а также ядро глубокого анализа на основе мультимодального трансформера. Модель интегрирует эмбеддинги изображений (ViT), текстовых описаний (BioClinicalBERT) и табличных признаков, применяя механизмы внимания для формирования агрегированного представления случая пациента и выявления комплексных клинически значимых паттернов.

3. Создан работоспособный программный прототип на стеке технологий Apache Spark, Docker, PyTorch и Petastorm, демонстрирующий полный цикл работы платформы – от приёма и подготовки мультимодальных данных до генерации аналитических отчётов и клинических сводок. Прототип подтвердил возможность развёртывания системы как контейнеризированного решения, адаптированного к работе в защищённом ИТ-контуре медицинской организации.

4. Экспериментально подтверждена эффективность подхода: на синтезированных мультимодальных данных система продемонстрировала способность выявлять статистически значимые и клинически релевантные взаимосвязи (например, ассоциацию пневмонии с ХОБЛ) при высокой производительности обработки (полный цикл анализа 700 записей занял менее 15 секунд). Результаты подтверждают масштабируемость архитектуры и её потенциал для работы с большими историческими массивами данных.

Научная новизна исследования заключается в системном архитектурном решении задачи построения сквозной аналитической платформы для медицинских данных, которое формализует переход от задач консолидации и хранения к задачам глубокого кросс-модального анализа с использованием гибридных вычислительных моделей. Предложен целостный подход, сочетающий методологию распределённой обработки, передовые модели глубокого обучения для мультимодальных данных и принципы промышленной разработки (контейнеризация, модульность).

Практическая значимость определяется созданием прототипа, который соответствует стратегическим ориентирам цифровой трансформации здравоохранения РФ и может служить основой для построения клинических систем поддержки принятия решений (CDSS) нового поколения в многопрофильных стационарах и диагностических центрах

Перспективы дальнейших исследований связаны с преодолением выявленных ограничений и развитием функциональности платформы: валидация на реальных клинических данных, интеграция дополнительных модальностей (МРТ, ПЭТ, потоковые данные), углубление методов объяснимого искусственного интеллекта (XAI) для повышения доверия клиницистов, а также адаптация архитектуры для работы в режиме, близком к реальному времени.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Hao Y., Cheng Ch., Li J., et al. Multimodal Integration in Health Care: Development with Applications in Disease Management. *Journal of Medical Internet Research.* 2025;27. https://doi.org/10.2196/76557
2. Liu C., Ye F. A Review of Multimodal Medical Data Fusion Techniques for Personalized Medicine. In: *IC-BIS '25: Proceedings of the 4th International Conference on Biomedical and Intelligent Systems, 11–13 April 2025, Bologna, Italy.* New York: Association for Computing Machinery; 2025. P. 338–347. https://doi.org/10.1145/3745034.3745088
3. Krones F., Marikkar U., Parsons G., Szmul A., Mahdi A. Review of Multimodal Machine Learning Approaches in Healthcare. *Information Fusion.* 2025;114. https://doi.org/10.1016/j.inffus.2024.102690
4. Xie Ch., Ningc Z., Guo T., et al. Multimodal Data Integration for Biologically-Relevant Artificial Intelligence to Guide Adjuvant Chemotherapy in Stage II Colorectal Cancer. *eBioMedicine.* 2025;117. https://doi.org/10.1016/j.ebiom.2025.105789
5. Heydari M., Sarshar R., Soltanshahi M.A. Distributed Record Linkage in Healthcare Data with Apache Spark. arXiv. URL: https://arxiv.org/abs/2404.07939 [Accessed 21st November 2025].

6. Deshpande P., Rasin A., Tchoua R. Biomedical Heterogeneous Data Categorization and Schema Mapping Toward Data Integration. *Frontiers in Big Data.* 2023;6. https://doi.org/10.3389/fdata.2023.1173038
7. Acosta J.N., Falcone G.J., Rajpurkar P., Topol E.J. Multimodal Biomedical AI. *Nature Medicine.* 2022;28(9):1773–1784. https://doi.org/10.1038/s41591-022-01981-2
8. Musik S., Sasin-Kurowska J., Panczyk M. Bridging the Past and Future of Clinical Data Management: The Transformative Impact of Artificial Intelligence. *Open Access Journal of Clinical Trials.* 2025;17:15–33. https://doi.org/10.2147/OAJCT.S509921
9. Hagan N.K.A., Talburt J.R. SparkDWM: A Scalable Design of a Data Washing Machine Using Apache Spark. *Frontiers in Big Data.* 2024;7. https://doi.org/10.3389/fdata.2024.1446071
10. Valo P., Tran A., Baranton E., Haas H., Freyssinet E., Vrzáková H. Clinical Data Integration and Processing Challenges in Healthcare Caused by Contemporary Software Design. *Digital Health.* 2025;11. https://doi.org/10.1177/20552076251374233
11. Shrotriya L., Sharma K., Parashar D., Mishra K., Singh Rawat S., Pagare H. Apache Spark in Healthcare: Advancing Data-Driven Innovations and Better Patient Care. *International Journal of Advanced Computer Science and Applications.* 2023;14(6):608–616. https://doi.org/10.14569/IJACSA.2023.0140665
12. Tu Y., Lu Y., Chen G., Zhao J., Yi F. Architecture Design of Distributed Medical Big Data Platform Based on Spark. In: *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 24–26 May 2019, Chongqing, China.* IEEE; 2019. P. 682–685. https://doi.org/10.1109/ITAIC.2019.8785620

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Пожарский Роман Витальевич, аспирант, Воронежский институт высоких технологий, Воронеж, Российская Федерация.
e-mail: pozharskij2013@mail.ru

Roman V. Pozharsky, Postgraduate, Voronezh Institute of High Technologies, Voronezh, the Russian Federation.

Петрова Елена Сергеевна, старший преподаватель, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: lenoks.sokolova@mail.ru

Elena S. Petrova, Senior Lecturer, Voronezh State Technical University, Voronezh, the Russian Federation.

Статья поступила в редакцию 27.11.2025; одобрена после рецензирования 19.12.2025; принята к публикации 25.12.2025.

The article was submitted 27.11.2025; approved after reviewing 19.12.2025; accepted for publication 25.12.2025.